ORIGINAL

AT&T

Frank S. Simone
Government Affairs Director

Suite 1000
1120 20th Street, N.W.
Washington, DC 20036
202 457-2321
FAX 202 457-2545
EMAIL fsimone@att.com

EX PARTE OR LATE FILED

July 13, 1999

RECEIVED

JUL 13 1999

FEDERAL COMMUNICATIONS COMMISSION
OFFICE OF THE SECRETARY

Ms. Magalie Roman Salas, Secretary
Federal Communications Commission
445 Twelfth Street, S. W. – Room TWB-204
Washington, D. C. 20554

Re: *Ex parte*, CC Docket No. 98-56, Performance Measurements and Reporting Requirements for Operations Support Systems, Interconnection, and Operator Services and Directory Assistance

Dear Ms. Roman Salas:

On April 12, 1999 the Commission staff requested information on the use of statistical methodologies for evaluating an incumbent local exchange carrier's performance in the provisioning of operations support systems ("OSS") to requesting carriers. Specifically, the staff posed a series of questions regarding the statistical methodologies proposed by the Local Competition Users Group ("LCUG") and BellSouth. Attached please find AT&T's responses to the Commission staff's list of questions. Please file a copy of this Notice in the record of the above-captioned proceeding.

Two copies of this Notice are being submitted to the Secretary of the FCC in accordance with Section 1.1206 of the Commission's rules.

Sincerely,

ATTACHMENT

cc: C. Mattey
M. Pryor
J. Jennings
D. Shiman
C. Pabo
A. Kearney
E. Einhorn
J. Stanley

No. of Copies rec'd 0+2
List ABCDE

Federal Communications Commission
Washington, D.C. 20554

April 12, 1999

Mr. Frank S. Simone
Government Affairs Director
AT&T
Suite 1000
1120 20th Street, N.W.
Washington, DC 20036

Dear Mr. Simone:

In Appendix B of the Commission's Notice of Proposed Rulemaking regarding performance measurements and reporting requirements for operations support systems ("OSS"), interconnection, and operator and directory assistance (CC Docket No. 98-56), the Commission sought comment on the use of statistical methodologies for evaluating an incumbent local exchange carrier's ("ILEC's") performance in the provisioning of OSS to requesting carriers. A number of parties have proposed various statistical methods that the Commission could use in evaluating an ILEC's provisioning of OSS. As a result of reviewing the various proposals of statistical methodologies, the Bureau Staff has developed a list of questions concerning the proposals made by the Local Competition Users Group ("LCUG") and BellSouth, which are attached. Please provide written responses to these questions at your earliest convenience and file the responses in CC Docket No. 98-56.

If you have any questions, please contact Daniel Shiman at (202) 418-7153.

Sincerely,

Carol E. Mattey, Chief
Policy and Program Planning Division
Common Carrier Bureau
Federal Communications Commission

Questions Concerning the Statistical Methodology to Use for Evaluating Performance Measurements

Please note that this list of questions does not cover all of the issues that have been identified, or that may be considered important.

General Comparison

1. Please compare the BellSouth and LCUG proposed tests (and their proposed variants) according to the following criteria. Provide theoretical analysis (with mathematical proofs), data analysis, examples, and cites to relevant references, where possible.
   - Efficiency (power) of the test to detect discrimination, including higher variance of CLEC data
   - Ability to handle confounding variables
   - Ability to handle heteroscedasticity
   - Ability to handle correlation (dependency) of measures, and correlation of subcells for a measure generated by disaggregating according to confounding variables
   - Ability to handle nonnormality of the data and small sample sizes

Concerning Efficiency:

2. What is the relative power of the BellSouth and LCUG tests? Assuming disaggregation is done according to multiple confounding variables, which will create many disaggregated cells, please give OC (or power) curves (using identical assumptions for both tests) for the following alternative hypotheses (H1-H4):

   Null hypothesis: H0: no discrimination

   vs. Alternative hypotheses:

   H1: discrimination in all cells
   H2: discrimination in half the cells
   H3: discrimination in 25% of the cells
   H4: discrimination in just one cell

   Also examine the two tests' ability to detect discrimination AGAINST the CLEC under the following hypothesis:

   H3a: discrimination against the CLEC in 25% of cells, discrimination for the CLEC in another 25% of cells

   Are there any other alternative hypotheses that should be considered? Is there any reason to believe that one particular scenario of potential discrimination is most likely to occur or most important?

3. How important is it to balance the probability of Type I and Type II errors? Is there a mechanical formula that would adjust the critical values (and hence the probability of a Type I error) as the sample size varied? How can we explicitly measure the costs of a Type I and of a Type II error, as BellSouth suggests needs to be done?

Concerning Estimating the Variance:

4. Why is it desirable to use replication to estimate the variance? What advantages does this have over using an alternative method?

Concerning Aggregating the Data:

5. What are the specific advantages/disadvantages of using aggregation of the adjusted data (the BellSouth approach)? Compare to testing unadjusted aggregate data (LCUG's original proposal) and testing individual cells of disaggregated data (LCUG's recent approach)? In particular, consider the criteria discussed in question 1.

6. Are there tests that can be performed to determine the validity of the degree of aggregation that BellSouth proposes versus the degree of disaggregation LCUG proposes? Is there some middle ground that can be reached through such tests by aggregating some of the cells, where appropriate, and disaggregating where aggregation is not appropriate?

Concerning Dependency:

7. Isn't the replicate estimate of the variance also affected by dependency (i.e., correlation) in the data? This appears to be confirmed in Wolters. Which methodology is affected more by dependency in the data?

8. How much dependency is there in the data (between measures, wirecenters, over time)? How can this be determined? Should this be determined using statistical means, or by examining physical relationships involved between measures (dependence on common computer system or common cable), or by examining each event *ex post*? Can a covariance matrix be developed using weekly or daily data, or by matching the ILEC and CLEC data using the multiple cells created through disaggregation? How much will the dependency measured affect the probability of a Type I error for the LCUG method, and for the BellSouth method?

Concerning Normality of the Data, and Sample Sizes:

9. Are the data nonnormal? How can this be determined? What size sample do we need to get an approximately normal distribution of a mean? How can this be determined?

10. Is permutations testing the best way to handle small, nonnormal samples? What are the advantages and disadvantages of permutations testing? Are there any problems that small sample sizes create for BellSouth's proposed methodology? Could we see a comparison of the results using permutation testing with BellSouth's results?

Concerning Statistical vs. Competitive Significance of the Results:

11. Should a statistically significant difference in means be interpreted to mean that there is discrimination in the process? In other words, should we consider whether the observed

difference in means will have an economic impact on CLECs' business? Won't very large sample sizes tend to make even small differences in means statistically significant? How large should a difference in means be for a particular measure for it to be considered "competitively significant" and therefore discriminatory? How should this "threshold difference" be determined for each measure? How can a "threshold difference" be implemented for a testing procedure?

Concerning Confounding Factors:

12. Is it necessary to disaggregate according to every confounding factor? What are the advantages and disadvantages of doing so? Would it be possible to disaggregate only according to those confounding factors that are statistically determined to have an impact on the results?

AT&T's Responses to FCC's Questions
Dated April 12, 1999

General Comparison

1. Please compare the BellSouth and LCUG proposed tests (and their proposed variants) according to the following criteria. Provide theoretical analysis (with mathematical proofs), data analysis, examples, and cites to relevant references, where possible.
   - Efficiency (power) of the test to detect discrimination, including higher variance of CLEC data
   - Ability to handle confounding variables
   - Ability to handle heteroscedasticity
   - Ability to handle correlation (dependency) of measures, and correlation of subcells for a measure generated by disaggregating according to confounding variables
   - Ability to handle nonnormality of the data and small sample sizes

Response to Question No. 1:

The FCC has requested that AT&T compare "the BellSouth and LCUG proposed tests (and their proposed variants)." As a preliminary matter, it is important to clarify the various "BellSouth and LCUG proposed tests" at issue.

BellSouth's initial statistical methodology was based upon its statistical process control ("SPC") model. However, BellSouth's proposed SPC methodology was rejected by both the Florida and the Georgia Commissions as an inappropriate method to detect discriminatory performance.[1] Following the rejection of its proposed SPC

[1] Order No. PSC-97-1459-FOF-TL, issued November 19, 1997 in Consideration of BellSouth Telecommunications, Inc.'s Entry Into InterLATA Services Pursuant to Section 271 of the Federal Communications Act of 1996, Docket No. 960786-TL (Fla. Pub. Serv. Comm'n), p. 183; Order issued May 6, 1998 in In the Matter of Performance Metrics for Telecommunications, Interconnection, Unbundling and Resale, Docket No. 7892-U (Ga. Pub. Serv. Comm'n), p. 16.

methodology by the Georgia Commission, BellSouth abandoned that methodology.[2] BellSouth next proposed a "Replicate Variance" method that was slightly modified in its later "jackknife" statistical model. BellSouth's Replicate Variance method, as well as its jackknife model, is inherently flawed.

In various submissions before the FCC and the Louisiana Public Service Commission, BellSouth has compared its jackknife statistical model against the following three so-called "LCUG" tests: (1) a "modified z" test that BellSouth asserts is a "straightforward calculation of the LCUG statistics applied to the whole data set without any adjustments"; (2) the "Adjusted modified z1" test that BellSouth asserts uses a "weighted variance estimate" that is designed "to correct for bias in the numerator of the test statistic"; and (3) the "Adjusted Modified z2" test that BellSouth asserts "uses a weighted average of subclass estimates" that is designed to correct for bias. See "Error Probabilities of the BST-Adjusted Jackknife Test," p. 3 attached to Letter from Victoria K. McHenry to All Parties on the LPSC Official Service List, Docket U-22252, dated April 19, 1999. Despite BellSouth's assertions to the contrary, these three so-called "LCUG" tests have never been endorsed by LCUG. Indeed, these purported "LCUG" tests are nothing more than BellSouth's interpretations of three different applications of the LCUG z statistic.

---

[2] See Transcript of Stacy Testimony, May 18, 1998, in BellSouth Telecommunications, Inc.'s Entry Into Long Distance (InterLATA) Service in Tennessee Pursuant to Section 271 of the Telecommunications Act of 1996, Docket No. 97-00309.

The only statistical test that has been endorsed by LCUG is set forth in Local Competition Users Group, Statistical Tests for Local Service Parity, Version 1.0, February 6, 1998 ("LCUG Version 1.0").[3] In this regard, the modified z-statistic endorsed by LCUG[4] accounts for both differences in means and variations from the mean.[5] In order to create a single test that can account for differences in means and performance variability, LCUG proposed a modification to the z-statistic that will make the statistical test have some power to assess whether the ILEC variance in its performance for CLECs is greater than the variance in its performance for itself. Under the modified z-statistic, a calculation is based on the assumption that the variability of the process serving the ILEC is equal to the variability of the process serving the CLEC. Further, under the modified z-statistic formula adopted by LCUG, the variance used in

---

[3] LCUG Version 1.0 sets forth the calculations and rationale for the modified z-statistic.

[4] LCUG's modified z-statistic has been recommended by the Public Utility Commission of Texas and agreed to by SWBT and found by the Michigan Public Service Commission to be the "most useful" statistical methodology to "reveal excessive variability within the samples as well as excessive differences between the calculated means." See Opinion and Order issued May 27, 1999 in In the Matter of Ameritech Michigan's Submission on Performances Measures, Reporting, and Benchmarks, Case No. U-11654 (Mich. Pub. Serv. Comm'n), p. 7; Letter from Melanie S. Fannin to ALJ Katherine D. Farroba, Public Utility Commission of Texas, dated April 26, 1999, attaching SWBT's Memorandum of Understanding in Investigation of Southwestern Bell Telephone Company's Entry into Texas InterLATA Telecommunications Market, Project No. 1625 (Tex. Pub. Util. Comm'n).

[5] The FCC has requested that AT&T provide "cites to relevant references" to the various statistical tests at issue. The only published paper of which AT&T is aware that discusses the LCUG z statistic is C. Brownie, D. Boos, and J. Hughes-Oliver, "Modifying the + and ANOVA F Tests When Treatment is Expected to Increase Variability Relative to Controls," 46 Biometrics pp. 259-266 (1990). However, this paper discusses only the basic theory and power functions for a single test and does not address aggregation.

the denominator is the variance of the ILEC's performance for itself during the reporting period.

When AT&T endorsed LCUG Version 1.0, it assumed that the LCUG statistical test would be applied in such a manner that appropriate "apples-to-apples" or "like-to-like" comparisons of CLEC and ILEC data could be made. In that connection, BellSouth and AT&T agree that "[t]he need for like-to-like comparisons requires the data to be disaggregated to a very deep level." See Letter from Kathleen B. Levitz to the Hon. Magalie Roman Salas, May 20, 1999 ("5/20/99 Levitz Letter"). However, LCUG Version 1.0 does not address the appropriate level of disaggregation that is needed to facilitate such "like-to-like" comparisons. Indeed, that issue has been the subject of negotiations between AT&T and BellSouth.

In any event, each of the three so-called "LCUG" tests – BellSouth's three different versions of the LCUG z statistic -- suffers from certain infirmities. Alternatively, another statistical methodology – that has not been subjected to extensive analysis but should be considered – is a statistical test using truncated z-scores. Set forth below is a comparative analysis of: (1) BellSouth's Replicate Variance/ jackknife statistical method; (2) "BellSouth's Modified z" test; (3) the "Adjusted Modified z1" test; (4) the "Adjusted Modified z2" test; and (5) the "truncated" method.

(1) BellSouth's Replicate Variance/Jackknife Method.[6]

Efficiency to Detect Discrimination. BellSouth's original "Replicate Variance" method (that is slightly modified in its later "jackknife" method) is based upon a technique for variance estimate generally referred to as the random group or "replicated method." Under this methodology, BellSouth uses a Replicate Variance Estimate method to provide a scale on which to compare differences between BellSouth and CLEC means for each type of order within each wire center. For each wire center, BellSouth computes the difference between the CLEC mean and the adjusted BellSouth mean and adds the individual cell differences within wire-centers, weighted by CLEC sample size, to obtain wire-center averages. BellSouth next separates the data into 30 separate sets, or replicates, of wire-centers so that each set is purportedly evenly distributed. BellSouth then averages within these groups to obtain group averages and finally averages over groups to obtain an overall average. BellSouth finally compares this overall average with a measure of spread obtained from the spread of the group averages.

---

[6] BellSouth's jackknife method is a "subsample replication technique." See "Follow-On Statistical Analysis of BellSouth Telecommunications, Inc.", p. 4 attached to Letter from Kathleen B. Levitz to the Hon. Magalie Roman Salas dated March 3, 1999 ("3/3/99 Levitz Letter"). Under the replicate variance methodology, the replicate estimate is calculated by using the observations in replicate g only. By contrast, under the jackknife methodology, the replicate estimate is calculated from the observations by removing the g^{th} group. As a practical matter, however, there are no significant, substantive differences between the Replicate Variance and jackknife tests. Accordingly, for purposes of its analysis herein, AT&T treats the Replicate Variance and jackknife methods as essentially one statistical test.

BellSouth's Replicate Variance test, as well as the jackknife model, is fundamentally flawed because it aggregates results using an overall (weighted) sum of the differences of each disaggregation that averages performance to such a degree that BellSouth can effectively "cancel" out bad performance where it is experiencing competition with good performance in other areas. Systematic differences in performance between wire-centers may cancel out under BellSouth's jackknife model, as well as systematic differences between cells within wire-centers. For example, if BellSouth repairs resold DS3 loops faster for CLECs than it repairs those types of loops for itself and repairs resold residence POTS much slower for CLECs than it does for itself, BellSouth's methodology will allow the good performance to cancel out the bad performance. Thus, BellSouth could effectively conceal its poor POTS performance through this technique.

Further, the jackknife method ignores systematic variation of results. The BellSouth methodology assumes that all differences between wire-centers, as well as all differences between cells within wire-centers, are random; BellSouth does not allow for the possibility that there are systematic biases in some wire-centers or between different cells within wire-centers. BellSouth's methodology allows the differences between wire-centers and between cells within wire-centers to contribute to the measure of spread to which the overall average is compared. Thus, if there are systematic differences between wire-centers or between cells within wire-centers, these will be counted as random effects and will be allowed for in the overall measure of spread. Because BellSouth makes the unjustified and dangerous assumption that variations are random, its methodology makes it possible for BellSouth to engage in

targeted discrimination. Under BellSouth's methodology, discriminatory, targeted actions at the wire center level may be concealed because of "good" performance to other CLECs in other locations. The inherent flaws in BellSouth's statistical model would leave CLECs highly vulnerable to undetected discrimination in performance.

Based upon the results of a simulation procedure, BellSouth claims that its proposed methodology has the power to detect discriminatory performance.[7] Under BellSouth's simulation procedure (id.):

- 240 wire centers with 2 classes each [were] created.
- Sample sizes [were] randomly generated and randomly allocated across subclasses.
- All subclasses means and standard deviations [were] generated based on wire centers and class effects.
- Correlation coefficients [were] randomly generated for each wire center.
- Orders within a wire center [were] generated so that the subclass ILEC-CLEC mean differences [were] correlated within a wire center.

Further, under BellSouth's simulation procedure, all of the observations within a cell were shifted up or down by some random amount; accordingly, the cell averages were more dispersed relative to the variations within cells. BellSouth claims that its simulation results demonstrate that its proposed methodology is insensitive to such correlations; or, stated differently, when correlations were introduced between the observations within each cell, the Type I and Type II error probabilities held constant at

[7] See, e.g., 5/20/99 Levitz Letter with attachments.

approximately 5%.[8] However, BellSouth's results show nothing of the sort. Quite the contrary, BellSouth's own simulation results demonstrate that its proposed methodology completely fails to detect deviations that could be attributable to biases between cells.

In this regard, BellSouth's simulation results could be attributable to consistent differences between wire-centers; in some wire-centers the ILEC mean could be consistently higher than the CLEC mean, while in other wire-centers, the reverse could be true. Indeed, by making a small change in BellSouth's simulation code, it is possible to arrange shifts within cells to occur systematically over the cells. Thus, a small change in BellSouth's simulation code could be made so that Cell No. 1 always obtains the smallest shift, Cell No. 2, the next smallest and so forth. However, BellSouth's jackknife model would fail to detect this systematic variation of results.

Relying principally on a book authored by K. Wolter titled Introduction to Variance Estimation, BellSouth trumpets its proposed "jackknife" methodology as a widely used statistical test. See "Follow-On Statistical Analysis of BellSouth Telecommunications, Inc.", p. 4, attached to 3/3/99 Levitz Letter. BellSouth's reliance on this slender reed must fail.

Wolter's analysis deals with the application of the variance estimate to large sample surveys. In such large sample surveys: a set of Primary Units (e.g. U.S.

---

[8] BellSouth notes that the simulation results for its proposed methodology show that: when there is no correlation between the performance measures, the Type I error rate is 0.052; when there is "medium correlation" between the performance measures, the Type I error rate is 0.064; and when there is "high correlation" between the performance measures, the Type I error rate is 0.059. See Chart titled "Simulation Results – Type I Errors" attached to 5/20/99 Levitz Letter.

cities) is defined and a random sample of such Primary Units is drawn; and a set of Secondary Units (e.g. districts) is defined within each Primary Unit, and a random sample of Secondary Units is drawn. Further, under the variance estimate methodology, all variability between Primary Units and between Secondary Units within Primary Units is regarded as being random. While the variance estimate methodology is entirely appropriate in the survey context where the units are drawn randomly from fixed, finite populations of Primary and Secondary Units, that methodology is inappropriate in the present context.

Unlike the large sample surveys in which random samples are drawn from a given population, BellSouth's methodology relies upon data from all wire centers. Indeed, the wire-centers are not drawn randomly from some population; but rather are simply all of the wire-centers that exist. Similarly, the cells within the wire-centers are not a random sample from some super-population of cells. For these reasons, BellSouth's reliance on the variance estimate methodology applicable to large survey samples is misplaced.[9]

---

[9] In Wolter's analysis (Theorem 2.2.1), an unbiased estimate of the variance of the overall mean d-double-bar is used based upon the assumptions that: (i) all of the component d-bar's have the same mean; and (ii) they may have different variances. While these assumptions may be entirely appropriate in a survey context, they are inappropriate in the present context. Under the null hypothesis, all of the component d-bar's have zero means, but the statistical test employed must detect those instances in which some, but not necessarily, all have positive means. It is not appropriate to consider only those instances where every sub-group has the same nonzero mean.

Ability to Handle Confounding variables. BellSouth appears to advocate disaggregation by wire center, time of month, and disaggregation levels defined by the Louisiana Public Service Commission. AT&T advocates using the levels of disaggregation proposed by LCUG in SQM Version 7.0. See Exhibit 3 (setting forth levels of disaggregation applicable to product and activity). Because AT&T does not have access to BellSouth's proprietary data, it is impossible for AT&T to assess the effect that other confounding variables might have on performance results. In any event, no matter how finely the performance data are disaggregated, BellSouth's methodology has the ability to handle such confounding variables.

Ability to Handle Heteroscedasticity. BellSouth's methodology ignores heretoscedasticity within cells and examines only the variance of certain aggregated group averages.

Ability to Handle Correlation. At the outset, it is important to distinguish among four kinds of correlation: (1) correlation between measures; (2) correlation over time; (3) correlations between cells; and (4) correlations within cells.

Correlations Between Measures. Correlations between measures must be addressed under BellSouth's methodology or any other statistical methodology. In this regard, a number of performance measures are calculated based on the same basic data. For example, measurements on average completion intervals and the percentage of orders completed within a given interval are calculated from the same data set. The resulting correlation between such measures must be allowed for when considering the statistical tests simultaneously; otherwise, double counting may occur. All of the statistical methods at issue must address this problem. BellSouth's

methodology, as well as all other statistical methods, can appropriately handle correlations between measures by estimating the size of the effect by simulation or simple observation over time.[10]

Correlations Over Time. The presence of correlation cannot be inferred from a single observation; as a consequence, the examination of performance data covering one month cannot establish the existence of correlations over time. On the other hand, an examination of data over several months can reveal the existence of such correlations. BellSouth's methodology, as well as other statistical methodologies at issue, must take such correlations into account when comparing performance results over several months.

Correlations Between Cells. As noted above, the BellSouth methodology ignores correlations between cells or treats them as random occurrences. As a result, BellSouth's methodology can effectively mask discriminatory performance.

Correlations Within Cells. Correlations within cells may be caused by the "clustering" of observations. The archetypical example is a backhoe accident

[10] BellSouth asserts that there is general agreement that "[c]orrelation between the performance measures must be accounted for by aggregating over similar measures." See Chart Titled "Open Issues – Independence of Performance Measures" attached to 5/20/99 Levitz Letter. This statement is flatly wrong. AT&T has reached no such agreement with BellSouth. As a result of a joint off-line session conducted as part of the proceedings before the Louisiana Public Service Commission, Dr. Colin Mallows agreed that "[c]orrelation between the performance measures must be accounted for in aggregation over performance measures." This statement simply means that, if the statistical methodology is designed to aggregate different measures, it must also allow for their correlation. However, BellSouth has contorted this statement and now claims that Dr. Mallows has somehow agreed that the appropriate way to allow for correlation between similar measures is to aggregate over them. Dr. Mallows has made no such concession.

that causes multiple outages that are fixed all at once. Because BellSouth's methodology ignores variability within cells, it is insensitive to correlations within cells caused by the clustering of observations.

Ability to Handle Nonnormality. BellSouth's methodology is largely impervious to nonnormality. In this regard, under BellSouth's methodology, only averages over large numbers of observations are used, and these averages will be close to normal no matter how the basic observations are distributed.

(2) BellSouth's Modified Z Test.

Efficiency to Detect Discrimination. As noted above, "BellSouth's Modified z" test – a test that it incorrectly describes as a "straightforward calculation of the LCUG statistics applied to the whole data set without any adjustments" -- has never been adopted by LCUG. In any event, that test, as described by BellSouth, is not a valid statistical methodology to test whether BellSouth is providing parity of access to CLECs. Because BellSouth's Modified z test would apply the LCUG statistics to unadjusted aggregate data, it will necessarily produce biased results and may fail to detect a lack of parity. Further, if the mix of services differs in different cells, simply pooling over all of the cells can introduce an appearance of discrimination overall even when there is parity in each cell.

Confounding Variables. It is important to disaggregate the data to a fine level so that appropriate like-to-like comparisons of CLEC and ILEC data can be made. Any statistical methodology that ignores important confounding variables can produce biased results. Every statistical methodology that disaggregates far enough will avoid

this effect. However, BellSouth's Modified z test cannot handle confounding variables because it uses unadjusted, aggregated data.

Heteroscedasticity. The LCUG z statistic uses only the overall variance; as a result, BellSouth's Modified z test is insensitive to variability within cells.

Correlation.

Correlation Between Measures. BellSouth's Modified z test can handle correlations between measures by estimating the size of the effect by simulation or simple observation over time.

Correlations Over Time. As noted above, all statistical methodologies must take correlations over time into account when comparing results from several different months.

Correlations Between Cells. BellSouth's Modified z test does not allow for correlations between cells.

Correlations Within Cells. BellSouth's Modified z test is sensitive to correlations within cells caused by the clustering of observations.

Nonnormality. BellSouth's Modified z test is largely insensitive to nonnormality because it uses averages and variances computed from aggregated data.

(3) BellSouth's "Adjusted Z1" Procedure.

As noted above, BellSouth has compared its proposed methodology to what it refers to as an "Adjusted Z1" test – a methodology that has never been adopted by LCUG. Under the Adjusted Z1 test, the numerator is the overall weighted average of the differences,

$$\sum n_{2j}(\bar{x}_{1j}-\bar{x}_{2j})/n_2$$

The denominator is

$$s_{1w}\sqrt{c_1+1/n_2}$$

where

$$s^2_{1w}=\sum(n_{2j}/n_{1j})\sum(x_{iju}-\bar{x}_{1w})^2/(n_2-1)$$

and $$c_1=\sum n^2_{2j}/n_{1j}/n^2_2$$

Assume the null hypothesis, so that BellSouth and CLEC observations have the same distribution and assume that in the j-th cell the mean is μ_j, and the variance is σ^2_j. Then the numerator has a mean of zero and variance of

$$\frac{1}{n_2^2}\sum\sigma_j^2\left(n_{2j}+\frac{n_{2j}^2}{n_{1j}}\right) \qquad (*)$$

which reduces, if the σs are all equal, to $\sigma^2(c_1+1/n_2)$. However, under the same assumptions, the expectation of s^2_1w is

$$\frac{1}{n_2-1}\left(\sum n_{2j}(\mu_j-\bar{\mu}_2)^2\sum\sigma^2+(n_{2j}-n^2_{2j}/n_2n_{1j})\right) \qquad (**)$$

which does not equal (*) even if all μ means are equal. When the means are different, their differences contribute to (**) but not to (*). Thus, because the denominator of the Adjusted z1 test is not an estimate of the correct variance, this methodology is fundamentally flawed.

Efficiency. The Adjusted z1 test has all of the failings of the BellSouth methodology. Because the Adjusted z1 test allows differences between wire centers and between cells within wire centers to contribute to the measure of spread to which

the overall average is compared, systematic differences in performance between wire centers and between cells within wire centers may cancel out. In addition, the Adjusted z1 test ignores systematic variation of results.

Confounding Variables. As noted above, any statistical procedure that ignores a relevant confounding variable can produce biased results. The Adjusted z1 test can handle confounding variables by disaggregating suitably.

Heteroscedasticity. As stated above, the Adjusted z1 test has the same deficiencies as the BellSouth methodology. The Adjusted z1 test ignores heteroscedasticity within cells.

Correlation.

Correlations Between Measures. The Adjusted z1 test can appropriately handle correlations between measures by estimating the size of the effect by simulation or simple observation over time.

Correlations Over Time. As is true with any of the statistical tests at issue, the Adjusted z1 test must take correlations over time into account when comparing performance results from several months.

Correlations Between Cells. The Adjusted z1 test relies on an overall measure of variance and, accordingly, treats all effects as random. If there are correlations between cells, the overall variance will not be estimated correctly.

Correlations Within Cells. The Adjusted z1 test is sensitive to the effects of clustering of observations because the overall variance will not be correct.

Nonnormality. The Adjusted z1 test is slightly sensitive to nonnormality; however, this should not be a problem if there are a large number of cells. The

numerator is an average over all of the data and, accordingly, will be close to normal. The denominator is also an average over all of the data and, as a result, will be a precise, but incorrect estimate.

(4) BellSouth's "Adjusted Z2" Procedure.

BellSouth has compared its proposed methodology against the so-called "Adjusted Z2" test -- another methodology that has never been adopted by LCUG. Under the Adjusted Z2 test, the numerator is the weighted average of within-cell differences. The variance of this is (*) above. In this variant, BellSouth estimates this by

$$\frac{1}{n_2^2}\sum s_j^2\left(n_{2j}+\frac{n_{2j}^2}{n_{1j}}\right)$$

where s_j^2 is the BST variance within the j-th cell.

Efficiency. As explained in more detail in AT&T's response to Question No. 2, when there are deviations of both signs, with the deviations averaging to zero being spread over the BST groups, the Adjusted z2 test is ineffective in detecting discriminatory performance. This is because the aggregation in the numerator, as in the BellSouth methodology, allows cancellation to occur.

Confounding Variables. All statistical methodologies, including the Adjusted z2 test, are vulnerable to producing biased results if important confounding variables are ignored.

Heteroscedasticity. The Adjusted z2 test can handle heteroscedasticity within cells, but will do so inefficiently if the heteroscedasticity is extreme. In this regard,

when the heteroscedasticity is extreme, the large-variance cells will tend to swamp the small-variance cells. Indeed, there could be substantial discrimination in cells with small variances that would not be detected because of the addition of differences having large variances in the numerator and the contribution of large variance estimates in the denominator.

Correlation.

Correlations Between Measures. The Adjusted z2 test can handle correlations between measures by estimating the size of the effect by simulation or simple observation over time.

Correlations Over Time. As is true of all of the statistical methodologies at issue, the Adjusted z2 test must take correlations over time into account when comparing performance results from several months.

Correlations Between Cells. The Adjusted z2 test relies on a within-cell measure of variance and appropriately treats the effects as biases. To understand this issue, it is important to understand how such a correlation might be manifested. Assume that in some wire centers, discriminatory service is provided to the CLEC's business customers and preferential treatment is accorded the CLEC's residence customers; while, in other wire centers, the reverse is true. The effect of such conduct can be determined by computing a measure of discrimination for each type of customer in each wire center. Thus, for example, for each wire center, two measures -- a Business and a Residence measure -- could be used. A scatter plot of these Business and Residence pairs (with one point for each wire center) would show a correlation. It is then important to determine whether this effect should be treated as a

random occurrence, or as a systematic bias. The Adjusted z2 treats all such effects as being potential biases, rather than as random occurrences.

Correlation Within Cells. Because the Adjusted z2 test examines variability within cells, it is sensitive to correlations within cells attributable to the clustering of observations. If all observations occur in batches of size k, then the LCUG adjusted z-2 statistic is inflated by a factor of sqrt(k). In any event, it is important to determine whether the clustering effect actually occurs; and if so, the extent of its pervasiveness. If data are provided regarding the nature and extent of the clustering effect, the Adjusted z test can be adjusted (through modifications to sample size) so that it can appropriately handle correlations within cells caused by the clustering of observations. If only an average incidence of clustering exists, an "average" adjustment could be made by dividing the LCUG z by the square root of the average cluster size. Thus, if there are only a few cluster events, this average will be near 1, and the effect will be small. However, if a large proportion of the observations are in clusters, the effect can be large.

Nonnormality. If there are a substantial number of cells, the Adjusted z2 test will not be affected by the nonnormality of the data and small sample sizes. The Adjusted z test can be made completely impervious to nonnormality by using the permutation method to compute the individual z's for measured variables. For counted variables, the sensitivity of the Adjusted z2 test to nonnormality could be alleviated by using a transform method. Under this method, each proportion k/n is transformed into a number in the range (0,pi/2) using what is referred to as the arc sine transformation: x=arc sin(sqrt (k/n)). The effect of the transform method is to make the variances of the

transformed variable x very nearly constant (=1/4n) no matter what the true proportion is. The transform method produces results that are very nearly the same as those when using the permutation method, that, for counted variables, is simply a matter of computing a hyperegreometric tail probability and converting that number to a z-score.

(5) The Truncated Method.

Efficiency. Another statistical methodology, that has not been subject to rigorous testing but should be considered, is the "truncated" method. In testing for discriminatory behavior, it is important to disaggregate the observations in each service quality measurement into cells that are small enough so that "like" comparisons within each cell can be made. For example, in its analysis of Order Completion Interval, BellSouth disaggregated the data by wire-center and classified the data according to the following five attributes:

Order Type (Change, New, Transfer)
Dispatch/Non-Dispatch
Residence/Business/Special
Number of circuits (<=10, >10)
Half of month.

Under BellSouth's methodology, there are 72 cells within each wire-center. However, not all the cells will be occupied; and many cells will have only a small number of observations.

Once the data have been disaggregated and a statistic has been computed within each occupied cell, the data must be reaggregated in some way to obtain a single overall measure for this service quality measurement. Simply pooling over cells can generate biased results. For example, assume there are the following

counts of Non-Missed and Missed Repairs Appointments for Residence and Business customers for the CLEC and ILEC:

	Business		Residence	
	CLEC	ILEC	CLEC	ILEC
Non-Missed	90	900	80	7200
Missed	10	100	20	1800
Percent Missed	10.0	10.0	20.0	20.0

Because the percentages of missed appointments for both CLEC and ILEC Business and Residential customers are equal within each cell, there is no evidence of discriminatory performance. By contrast, if the data are aggregated by simply pooling these two cells, the resulting data show that the CLEC is accorded preferential treatment:

	CLEC	ILEC
Non-Missed	170	8100
Missed	30	1900
Percent Missed	15.0	19.0

To avoid this effect, as an initial step, it is necessary to calculate the differential performance for each cell in which there are both ILEC and CLEC observations. BellSouth simply calculates the difference between the ILEC mean and the CLEC mean for each cell and ignores the dispersion of the observations within each cell. In the above example, BellSouth would calculate a difference of zero between the two percentages within each cell, and the average of these (weighted by CLEC sample size) would be zero (as it should be). The LCUG formula can be applied to obtain a Z-score for each cell. Whenever a cell has only a small number of observations, a permutation calculation is required. In the example, the Z's would be zero.

Individual cell measures must be combined to obtain an overall index for this service quality measurement. The method used should have the following properties:

(a) The method should provide a single overall index that is on a standard scale;

(b) If the entries in the cells are exactly proportional, the reaggregated index should be very nearly the same as if the disaggregation process had not been used;

(c) The contribution of each cell should depend on the number of observations in it;

(d) As far as possible, cancellation should not be allowed to occur; and

(e) The index should be a continuous function of the observations.

Under the truncated method, the Z-scores are replaced by truncated Z-scores:

$$Z^{+} = \min(Z,0)$$

Positive values are replaced by zero. To obtain a weighted average of these truncated Z's, the following weight should be used:

$$\text{weight} = 1/\sqrt{\left(\frac{1}{nILEC} + \frac{1}{nCLEC}\right)}$$

Finally, this average should be adjusted to allow for the way it has been constructed:

$$Z^{*}_{agg} = (\text{sum } (WZ^{*}) - M)/\sqrt{V}$$

where M and V are the theoretical mean and variance of sum (WZ^+), assuming perfect parity everywhere. Calculating M and V is straightforward, once the sample sizes are known. Under the truncated method: a single overall index is provided on a standard scale; a reaggregated index is generated that would be very nearly the same even if the disaggregation process had not been used; the contribution of each cell will depend upon the number of observations in it; cancellation will not occur; and the index will be a continuous function of the observations. Exhibit 1 refers to certain minor technical difficulties that should be addressed in the calculation of M and V when the sample sizes are small.

Confounding Variables. As is the case with all statistical methodologies, the truncated method will produce biased results if an important confounding variable is ignored.

Heteroscedasticity. Since the truncated method calculates z's within cells, the method is unaffected by heteroscedasticity within cells.

Correlation.

Correlation Between Measures. A number of performance metrics within the same service quality measurement categories are calculated from a common set of the same basic data. Thus, for example, a measurement of the average completion interval and distributional measures such as the percentage of orders completed within a given interval are calculated from a common set of data. The resulting correlation between the measures must be allowed for when statistical tests are employed simultaneously. All of the statistical methods at issue face and must deal with this issue. Each statistical methodology at issue, including AT&T's proposed

truncated method, can address the effect of correlation between measures by estimating the size of the effect by simulation or simple observation over time.

Correlations Over Time. As is true with all statistical methodologies, the truncated method must take into account all correlations over time when comparing performance results.

Correlations Between Cells. The truncated method handles correlations between cells by treating them as biases.

Correlations Within Cells. Because the truncated method examines variability within cells, it is sensitive to "backhoe" clustering -- correlations within cells due to the clustering of observations. However, if data are available revealing the nature and scope of the clustering effect, the truncated method can be adjusted, via modifications to sample size, so that it can properly handle the effect of clustering of observations.

Nonnormality. AT&T's proposed truncated method can appropriately handle nonnormality of the data and small sample sizes through permutation testing.

Concerning Efficiency:

2. What is the relative power of the BellSouth and LCUG tests? Assuming disaggregation is done according to multiple confounding variables, which will create many disaggregated cells, please give OC (or power) curves (using identical assumptions for both tests) for the following alternative hypotheses (H1-H4):

   Null hypothesis: H0: no discrimination

   vs. Alternative hypotheses:

   H1: discrimination in all cells
   H2: discrimination in half the cells
   H3: discrimination in 25% of the cells
   H4: discrimination in just one cell

Also examine the two tests' ability to detect discrimination AGAINST the CLEC under the following hypothesis:

H3a: discrimination against the CLEC in 25% of cells,
discrimination for the CLEC in another 25% of cells

Are there any other alternative hypotheses that should be considered? Is there any reason to believe that one particular scenario of potential discrimination is most likely to occur or most important?

Response to Question No. 2:

The attached graphs (Exhibit 2) show power functions for three methods of analysis (i.e. BellSouth's Replicate Variance method, the Adjusted z2 test, and the truncated method) for six kinds of deviations from the null hypothesis. In every case it is assumed that N=240 cells, each with a large number of observations for both BST and CLEC. In each cell it is assumed that the average difference between BST and CLEC observations is normal with mean mu_cell (depending on the hypothesis being studied) and variance 1. It is also assumed that sample sizes are so large that the variability in the sample variances can be ignored.

BellSouth's methodology is applied using G = 30 groups; so there are 8 cells in each group. In this method the CLEC-BST differences within each group are averaged, and the average of these mean differences is divided by the square root of their variance. This produces a statistic

$$BST = \text{average}(m\_g) / \text{sqrt}(\text{variance}(m\_g)/G)$$

which has (very nearly) a standard normal distribution under the null hypothesis.

Under the Adjusted z2 test, z-scores that are obtained for each cell are simply averaged over the cells. This produces the standardized statistic

$$aveZ = \text{sqrt}(N)\ \text{average}(Z\_cell)$$

By contrast, under the truncated method, the z-scores are truncated by replacing positive values with zero. Thus $Z^* = \min(Z,0)$. The average of these truncated Z's is then adjusted to compensate for the truncation:

$$\text{truncZ} = \text{sqrt}(N)\ (\text{average})Z^*\_\text{cell}) - a)/b$$

where $a = -1/\text{sqrt}(2^*\text{pi})$, $b^2 = ½ - 1/(2^*\text{pi})$.

These three methods were applied to eight kinds of departures from the null hypothesis. In each case, the total absolute shift is $N^*\text{mu}$. The alternatives are:

1. "allshift": in each of the N cells, the difference between the CLEC-BST data has the same mean mu > 0;
2. "halfshift": in half of the cells, the difference has mean 2^*mu; in the other half the mean is zero. There are two variants:

   21: the non-zero means are spread evenly over the 30 BST groups;

   22: the non-zero means are concentrated in half of the 30 groups.

3. "quartershift": in one quarter of the cells, the difference has mean 4^*mu; in the remaining 3/4 the mean is zero. Again, there are two variants:

   31: spread (as evenly as possible) over the BST groups;

   32: concentrated in 1/4 of the groups.

4. "oneshift": in one cell the mean difference is $N^*\text{mu}$; in the remaining cells, the mean difference is zero.